

DOCUMENT RESUME

ED 464 973

UD 034 987

AUTHOR Berlak, Harold
TITLE Academic Achievement, Race, and Reform: Six Essays on Understanding Assessment Policy, Standardized Achievement Tests, and Anti-Racist Alternatives.
PUB DATE 2001-11-00
NOTE 35p.
PUB TYPE Collected Works - General (020) -- Opinion Papers (120)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Diversity (Student); Educational Change; Elementary Secondary Education; *Equal Education; Higher Education; *Minority Group Children; Multicultural Education; Public Schools; Standardized Tests; Student Evaluation; *Test Bias
IDENTIFIERS *Achievement Gap; *Social Justice

ABSTRACT

This set of six essays was written as a resource for those working in their schools and communities to promote social justice, combat racism, and encourage quality education for all youth. The six essays address the following set of questions: (1) What is the Academic Performance Index (API)? Where did it come from? What are the connections between Goals 2000, Bush's educational policies, the API, and the Stanford 9 Achievement Test? (2) What explains the achievement gap between white students and students of color? (3) What is the impact of standardized testing on diversity in the curriculum? (4) What are normed tests and are they valid measures of academic performance? Are normed tests racist? Do criterion-referenced tests remedy the limitations of normed tests? (5) What are the consequences of centralizing control over the assessment process?; and (6) What are the obstructions to change and the practical alternatives to the current policies? (SM)

Academic Achievement, Race, and Reform

Six essays on understanding assessment policy, standardized achievement tests, and anti-racist alternatives

Harold Berlak

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

✓ This document has been reproduced as
received from the person or organization
originating it

□ Minor changes have been made to
improve reproduction quality

• Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL HAS
BEEN GRANTED BY

H. Berlak

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

Introduction

This set of essays was written to serve as a resource for those who are working in their schools and communities for social justice, to combat racism, and promote quality education for all youth.

The essays address the following set of questions:

1. What is the Academic Performance Index or API? Where did it come from? What are the connections between Goals 2000, Bush's educational policies, the API and the Stanford 9 Achievement Test?
2. What explains of the 'achievement gap' between white students and students of color?
3. What is the impact of standardized testing on diversity in the curriculum?
4. What are normed tests and are they valid measures of academic performance? Are normed tests racist? Do 'criterion-referenced' tests remedy the limitations of normed tests?
5. What are the consequences of centralizing control over the assessment process?
6. What are the obstructions to change, and what are the practical alternatives to the current policies?

While there is overlap, each essay focuses on one set of questions and can be read independently, and in any order. I do not attempt to cover all the important questions and issues. My intent in each essay is to foreground the issue of race and institutional racism, which often remains in the shadows in discussions of testing and assessment policy.

Harold Berlak

June, 2001 (updated November, 2001)

Overview and Origins of Current Assessment Policies

Case in point: California's Academic Performance Index or API

In January, 2000 California installed a statewide public school ranking system called the Academic Performance Index or API. The State Department of Education released to the public and posted on its website comparative rankings of every public school in the state. The rankings made the headlines in all the State's major newspapers, and was a lead story on the nightly local TV news. Newspapers were filled with pages of tables and graphs comparing scores of local schools and districts across the state.

What the API and all indexing systems provide is a uniform scale for measuring educational productivity statewide, what one editorial writer approvingly called the equivalent of a Dow Jones Average for schools. Every public school in California is given a number from 200-1000 based on the school's average score on the Stanford 9 Achievement Test or *Stat-9*, a standardized test published, distributed, and scored by Harcourt Educational Measurement. Every school is also categorized according to the relative affluence of the area it serves, and within its category ranked 1-10, worst to best based on the school's average on the *Stat-9*. At some point the high school exit exam scores and other factors may enter into the calculation of API, but for now and for the next several years performance on this single test, the *Stat-9*, is by state regulation the final and controlling measure of students' academic performance, school quality, and the professional competence of teachers and, principals and school superintendents.

As of this writing most states have installed or are in the process of installing an *indexing policy* that ties test performance to a system of rewards and sanctions. An indexing system consolidates state control by linking test results to specific sanctions and rewards, and by creating an annual high profile public display of test results. A uniform and unitary measure of educational productivity also sets the stage for market driven solutions to educational problems with not-for-profit and profit-making educational management organizations competing with the public education sector to produce high test scores.

Three features are common to all existing and proposed "high stakes" indexing policies:

- (1) School districts are required to administer the same test to all students with no (or almost no) exceptions.

- (2) There is sole or very heavy reliance on standardized tests as the measure of 'academic achievement', either a 'norm-referenced' or 'criterion-referenced proficiency' test in one or more academic school subjects.
- (3) Test performance is linked to a centralized state system of individual and institutional rewards and sanctions.

Its supporters claim that high stakes indexing systems not only raise standards, but also promote educational opportunity because the process is color blind and objective, thus free of racial and cultural bias. California's API, for example, sets a score of 800 as the standard of excellence for all schools regardless of the socio-economic class, race, or languages of the communities the school serves. Despite great disparities in physical plant and human and material resources between rich and poor schools, proponents claim the system is fair because in measuring a school's progress toward meeting 800 standard and in distributing rewards, richer schools and poorer schools are compared to their own kind.

Schools whose average scores fall below the 800 mark are considered below standard and failing. The school district, principal and school staff are put on notice by the State. Should the school repeatedly fall short of the targets, it will lose state and federal funding, and could be taken over by the state. What this means is not yet entirely clear, but is generally taken to include 'restructuring' and 'reconstituting' failed schools. The State of California's Department of Education could assume direct control, or contract with a non-profit or a profit-making management company. The entire staff of the failed school, principals, teachers, and counselors, could be reassigned, demoted or, if without tenure, fired. Though not formally a part of California's API policy, there are dire consequences for individual students who fall below the established norms or cut scores on the mandated tests. They could be required to repeat a grade, placed in an 'opportunity' or remedial track, barred from entry to special academic programs and high status public schools, or denied a high school diploma.

In addition to the sticks, there are carrots. Schools, students, and teachers meeting API targets are rewarded with access to educational programs, opportunities for professional advancement, scholarships and cash. In 2001, 12,250 teachers in low scoring schools that beat their previous year's scores by an average of 5%-10% or more are in line to receive bonuses of five to twenty-five thousand dollars. Both the schools and the teachers within schools that decrease the shortfall to the 800 mark by 5% or more are also eligible for monetary rewards. The actual amount depends on annual state educational appropriations and numbers of schools that qualify. In 2001 the reward is approximately \$60 per student or \$60,000 per thousand students, a considerable sum. Teachers and all school staff are also to receive about \$600. In addition, highest scoring students on the Stat-9 were rewarded with \$1000 scholarships.

'High stakes testing' is the term most often when tests are use for making fateful and often irreversible educational decisions. It is important to note that what marks Stat-9 or any test as " high stakes" is not the test itself, but the fact that the government

authorities use test scores as the basis for distributing rewards, and punishments to teachers, students, and school administrators.

Goals 2000 and Indexing Policy.

Both state mandated standardized testing, and indexing policies are a product of a larger educational standards and accountability agenda that marches under the banner of 'Goals 2000.' It is a policy first announced by George H. Bush in his inaugural address,¹ pursued by Clinton, and actively supported by George W. Bush.²

The policy simply stated equates high academic standards with high test scores. Goals 2000, a piece of legislation passed by Congress in 1994, proposed and signed into law by Bill Clinton, made six lofty promises to be met by the year 2000³ including achieving 'world class' academic standards. Clinton's original proposal contained provisions for national testing in reading and math. The testing provision was eliminated because of organized opposition by the left and right margins of American electoral politics, uniting civil rights, children's and fair test advocates with Christian right wing and libertarian Republicans. The opposition was led in the House by the Black Caucus and in the Senate by then-Senator John Ashcroft. What remained in the bill were a variety of federal government incentives to induce state governments to assume a higher degree of state control over the schooling process by installing statewide testing.

Though Goals 2000 policies and its political and its effects are widely reported in the education press, it went almost unnoticed in the national and local media and journals of opinion, right, left, and center. In March 1994, shortly after Congress passed Goals 2000, an education writer for the NY Times called the shift of control in the bill historic, and unprecedented, yet the story itself was buried in the inside pages. The public's and the press' indifference to Goals 2000 was, however, understandable. To most, it appeared harmless, did not raise taxes, and had no noticeable immediate effects. Life in schools went on more or less as before. Impact on schools and the everyday lives of teachers, school administrators, parents, and students would be enormous but was several years away.

1 *Raising Standards For American Education. A Report to Congress, the Secretary of Education, the National Goals Panel, and the American People*; Wash DC 1/24/1992 US Printing Office ISBN 0-16-036097-8. This report explicitly acknowledges and adopts California as the model. The 'smart' standards tied to 'smart' test policy was introduced to California by Bill Honig nominally a liberal Democrat who served as State Superintendent of Instruction from 1983-93.

3 These goals are: by the year 2000 (1) all children will start school ready to learn; (2) the high school graduation rate will increase to at least 90%; (3) all children will leave grades 4, 8, and 12 having demonstrated competency in challenging subjects including English, mathematics, science, foreign languages, civics and government, arts, history and geography. (4) American students will be first in the world in science and mathematics achievement; (5) every adult American will be literate; (6) every school will be free of drugs, violence, and the unauthorized presence of fire arms and alcohol and will offer a disciplined and drug free environment.

In January 2000, the full weight of Goals 2000 first came to California in the guise of the API. The API is Goals 2000 in action with one very notable departure from the original design as espoused in government reports, and officials in testimony before Congress. The original Goals 2000 design called for 'world class standards' tied to a new breed of 'world class' tests that were to be capable of assessing in-depth understandings, complex ideas and 'higher order' thinking required by the new demanding standards. These were to be smart tests or 'tests worth teaching to' a phrase coined by Goals 2000 advocates. Although there are some continuing efforts to create smart tests tied to smart standards, they remain exceptions to the rule.

California spent about 100 million-dollar to produce the first ever state mandated smart test. It was called the CLAS language test and was 'aligned' to the State's 1987 standards. This set of standards (called a 'curriculum framework' in California) sought to reform the way reading and writing were taught. Contrary to the critics' claims, this framework did not abandon phonics. It did, however, represent a departure from sole reliance on commercially published 'basal' readers that were tied to workbooks, and it encouraged literature-based basic reading instruction, and a writing curriculum that emphasized making connections to the learner's interests and experience.

What was sold to the public as an apolitical, non-partisan plan became deeply mired in California's cultural wars and toxic electoral politics. Educational Testing Service developed the first of the new tests called CLAS which was aligned to the State's new language standards. It arrived in 1994 on the eve the mid-term election that gave Republicans control of the House and Senate for the first time in forty years and elevated Newt Gingrich to Speaker of the House.

This was the same election that gave Pete Wilson a second term as governor. Wilson considered a 'moderate' Republican because of his pro-choice view was unpopular and entered the campaign with low voter approval in part because California's economy was sluggish in the early 1990's, and in part because the politically powerful Southern California Christian right-wing Republicans were indifferent, if not hostile to Wilson because of his stance on abortion.

Wilson won a second term in a close election by riding the tails of overwhelming white male voter support for two infamous propositions, 185 and 187. Proposition 185 called the 'three strikes you're out' initiative mandated 25 year to life sentences for a third felony regardless of its seriousness. 187 sought to deny schooling and virtually all social and health services to undocumented immigrants, the vast majority of whom are from Mexico. Wilson pandered to the Christian right which had bitterly opposed adoption of the new language standards. He made CLAS and the language standards into a hot button, racially charged issue, saying that they served the ideological aims of the left-wing and of multicultural extremists. Among the offensive test items cited by Wilson as evidence of political correctness and multiculturalism run amuck was a passage taken from Maxine Hong Kingston's *Women Warrior*, followed by the directions, 'Write an essay in which you interpret the moments of silence or inability to speak.'

The progressive and moderate forces that had supported the policy of 'smart' tests tied to new 'smart' standards were outmaneuvered politically, overwhelmed, and soundly defeated. In the waning days of the Wilson Administration, the legislature passed a bill creating the Academic Performance Index or API. In 1998 CLAS was replaced with the *Stat-9*, a run-of the-mill commercially available standardized, multiple-choice achievement test published by Harcourt Measurement. In 2000, the *Stat-9* became the state's instrument for calculating the API.

2 Race and the Achievement Gap

That there is a race gap in educational achievement is not news. Large numbers of the nation's children leave school, with and without high school diplomas, barely able to read, write, and do simple math. But the failures of the schools are not evenly distributed. They fall disproportionately on students of color.⁴

Even when parents' income and wealth is comparable, African-Americans, Native Americans, Latinos, and immigrants for whom English is not a first language lag behind English-speaking, native born, white students. The evidence for the gap has been documented repeatedly by the usual measures. These include drop-out rates, relative numbers of students who take the advanced placement examination, who are enrolled in the top academic and 'gifted' classes and/or admitted to higher-status secondary schools, colleges, graduate and professional programs. And last but certainly not least, are the discrepancies in scores on standardized tests of academic achievement, on which teachers' and students' fate so heavily depend.

How is this achievement gap to be explained? This essay focuses first on the general question and then separately on the statistical gap in standardized test scores. It then discusses the crucial distinction between academic performance and academic achievement as measured by standardized tests. Though often spoken of as though they are one, they are clearly different. The failure to separate out the difference clouds and confounds educational and policy issues and misleads us in efforts to explain and eradicate the race gap in academic performance.

Explanations over the years

Over the years, the major reasons given for the claimed superior attainments of whites in cultural, artistic, and academic endeavors were overtly racist. It was said that the explanation lay in the superior genes of white northern European, Anglo-Americans. As the social sciences developed in the latter years of the 19th and the 20th centuries, 'scientific' tracts defending white supremacy appeared with regularity. By the 1930's, the eugenics movement (which posited a biological basis for the superiority of whites) managed to gain a foothold in North American universities. And, it is relevant to add, all the leaders of this overtly racist movement were the leaders of the newly emerging field of scientific mental measurement. Many were the same men who testified before Congress in the early 1920's and lent scientific credence to the racist immigration exclusion acts which barred or greatly restricted immigration from Asia, Latin America, and southern and eastern Europe. The eugenics movement was considered a respectable academic discipline until it was discredited following the defeat of the

⁴ *Still Separate, Still Unequal, A Research Brief* Oakland, CA: Applied Research Center, May, 2000. www.arc.org

Third Reich and the immensity of the crimes committed in the name of Nordic racial purity.⁵

In 1969, the scientific case for racism was revived by an article published in the *Harvard Educational Review* by University of California-Berkeley education professor Arthur Jensen. Based on his statistical analysis of I.Q. test scores, he concluded that African Americans were genetically inferior to whites in general intelligence. His racist thesis was widely disseminated and discussed in the popular press and in respectable academic and policy circles. In time, Jensen's conclusions were thoroughly discredited by a spate of books and articles.⁶ In 1994, once again using standardized test data, Charles Murray and Richard Herrnstein in *The Bell Curve* claimed to have proven that the inferior place of black and brown people in the social, political, and economic order was rooted in biology. The arguments for the genetic superiority of the white race were again dismembered and discredited by many geneticists and biologists.⁷

Recently a more subtle form of 'scientific' racism has gained some respectability. The inferiority of the Black and brown races is now said to lie not necessarily in genetics but in culture and history. This more quietly spoken academic version of the master-race ideology has also been thoroughly dismantled, yet racist explanations for the race gap persist.⁸

Once all 'scientific' arguments supporting racism are dismissed how is the ever-present gap in academic school performance to be explained? Numerous social and behavioral scientists have addressed this question.

A statistical study by Professor Samuel Meyers Jr. at the Roy Wilkins Center for Human Relations and Social Justice at the University of Minnesota sought to determine whether poverty was a primary cause of the poor performance of Black students on the Minnesota Basic Standards Test.⁹ Passing this test was scheduled to become a prerequisite for a high school diploma in 2000. In a 1996 trial run in Minneapolis, 75 percent of African-American students failed the math test, and 79

⁵ See Stephen Selden *Inheriting Shame: The Story of Eugenics in America*. New York, Teachers College Press, 1999.

⁶ Steven J. Gould, 'Jensen's Last Stand,' *New York Review of Books*, 1980; Leon J. Kamin, *The Science and Politics of IQ* New York 1974; Daniel M. Kohl, 'The IQ Game: Bait and Switch,' *School Review* 84:44 1976., John Wiley

⁷ Russell Jacoby and Naomi Glauberman (eds.) *The Bell Curve Debate*. New York, NY Times Books/Random House, 1995

⁸ Cultural supremacy arguments are dismantled in Jared Diamond, *Guns, Germs and Steel* New York: Norton, 1997.

⁹ Samuel L. Myers Jr. and Cheryl Mandala *Is Poverty the Cause of Poor Performance of Black Students on Basic Standards Examination?* Roy Wilkins Center for Human Relations and Social Justice, Univ. of Minnesota. Paper presented at the 1998 American Educ. Research. Assoc. Annual Meeting, June, 1997

percent failed in reading, compared to 26 percent and 42 percent respectively for whites.

The researchers found that, contrary to expectations, test scores were *not* statistically related to school poverty, neighborhood poverty, racial concentration, or even ranking of schools (except in the case of whites). They did find that African Americans, American Indians, and Hispanics were underrepresented in the top ranked schools.¹⁰ African-Americans were 4.5 times as likely to be found in schools ranked low in math, and twice as likely to be found in schools ranked lowest in reading.

For both white and students of color, success on the tests was positively correlated to how an individual had been tracked. Only 6.9 percent of students of color compared to 23 percent of white students had access to 'gifted and talented' programs. This study suggests that tracking and the quality of the academic opportunities available in the school affects both the test score gap and the gap in academic performance generally. While these correlational studies are suggestive, they do not examine basic causes nor explain the pervasiveness and stability of the gap over prolonged periods of time.

A set of experimental studies conducted by Stanford University professor Claude Steele, an African-American psychologist, sought to explain the circumstances and situations that give rise to the race gap in test scores.¹¹ He and colleagues gave equal numbers of African-American and white Stanford sophomores a 30-minute standardized test composed of some of the more challenging items from the advanced Graduate Record Examination in literature. Steele notes all the students were highly successful students and test-takers since all Stanford students must earn SAT scores well above the national average in order to be admitted to the university.

The researchers told half the students that the test did not assess ability, but that the research was aimed at 'understanding the psychological factors involved in solving verbal problems.' The others were told that the test was a valid measure of academic ability and capacity. African American students who were told that the test was a true measure of ability scored significantly lower than the white students. The other African-American students' scores were equal to white students'. Whites performed the same in both situations.

The explanation Steele offers is that Black students know they are especially likely to be seen as having limited ability. Groups not stereotyped in this way do not experience this extra intimidation. He suggests that 'it is serious intimidation, implying as it does that if they should perform badly, they may not belong in walks of life where their tested abilities are important -walks of life in which they are heavily invested.' He labels this phenomenon 'stereotype vulnerability.'

¹⁰ Schools were ranked in terms of resources, education and experience of staff, number, depth and range of academic course offerings.

¹¹ Claude M. Steele, 'A threat in the air: How stereotypes shape the intellectual identities,' *American Psychologist*, 52, 1997. Also see 'Stereotyping and its threat are real' *American Psychologist*, 53, 1998.

In another study, Steele and colleagues found, to their surprise, that students most likely to do poorest on the tests were not the least able and prepared academically. To the contrary, they tended to be among the more highly motivated and academically focused.

While Steele's research provides a plausible psychological explanation for the gap, it does not probe the historical, social, and cultural factors that have created and continue to sustain these stereotypes. We are left with no explanation of how 'stereotype vulnerability' is created by and also shapes everyday life in society and at school.

The previously cited studies focus on the gap in standardized test scores. The next study cited is one of a large number of recent 'qualitative' studies - observational, historical, and ethnographic studies - that illuminate relationships of culture, gender, and race to the social relations within the classroom and school.¹²

Signithia Fordham, an African-American anthropologist, studied a Washington, D.C., public high school and focused on how the 'hidden' and explicit curriculum shapes student aspirations and achievements, and how students of differing cultural, racial, and social backgrounds respond to the schooling experience.¹³ Hers is a multifaceted, complex study, including interviews, participant observation, questionnaires, and field notes, gathered over a 4-year period. She concludes that for African-American students, patterns of academic success and underachievement are a reflection of processes of resistance that enable them to maintain their humanness in the face of a stigmatized racial identity. She shows that African-American adolescents' profound ambivalence about the value and possibility of school success is manifest as both conformity and avoidance. Ambivalence is manifest in students' motivation and interest in schoolwork, which of course includes mastery of standardized test-taking skills. The following two quotes are taken from interviews with two African-American men. The first is from a young lawyer employed in a Washington D.C. firm who had been a National Merit finalist and whose test scores were among the top 2% in his state.

Commenting on why he was disappointed with his career, he observed: 'I realized that no matter how smart I was [in school] or how hard I was willing to work [in the law firm] that it wasn't going to happen for me. ... Don't get me wrong, integration has been great for my life. Without it, I would be playing on a much more restrictive field, [but] there's no doubt in my mind that I would be much more successful today if I were white.'

¹² These include Lisa Delpit (1995) *Other Peoples Children* New York: The New Press . Joyce E. King, (1996) 'The Purpose of Schooling for African American Students' In J. King, E. Hollins and W.C. Hayman eds. *Preparing Teachers for Cultural Diversity* New York: Teachers College Press; Gloria Ladson Billings (1999) *The Dream Keepers: Successful Teachers of African-American Children*, San Francisco: Josey-Bass; Ann Berlak and Sekani Moyenda, (2001) *Taking it Personally: Racism in Schools*. Philadelphia: Temple. Univ. Press.

¹³ Fordham, Signithia, (1996) *Blacked Out, Dilemmas of Race, Identity, and Success at Capital High*, Chicago: University of Chicago Press.

A high-performing, African-American high school student offers the following view of why African Americans often underperform in school, and also expresses his doubts that his own school success will be rewarded.

'Well, we supposed to be stupid ... we perform poorly in school 'cause we all have it thought it up in our heads we're supposed to be dumb so we might as well go ahead and *be* dumb,' he said. 'And we think that most of the things we learn [at school] won't help us in life anyway. ... What good is a quadratic equation gonna do me if I'm picking up garbage cans?'

Fordham found that even the most academically talented African-American high school students expressed profound ambivalence toward schooling and uncertainty that they will reap the rewards of school success. Virtually all African Americans she interviewed indicated that a central problem facing them at school and in larger white society is the widely held perception by whites that African-Americans' are less able and intelligent and their continuing need to confront and deal with this reality in everyday experience.

These three studies taken together suggest three related explanations for the race gap in academic achievement and in test scores. First, is students' perceptions of the opportunities in the wider society and the realities of 'making it'. Second, are the educational opportunities available in the educational system itself -within school districts, schools, and within each classroom. Third, are the cumulative psychic and emotional effects of living in a social world saturated with racist ideology, and where racist practices and structures are pervasive, and often go unnamed.

Gap in test scores

There has been long-standing concern over the race gap in test scores. What is almost always overlooked is the size and educational significance of the test score gap. Most people assume that the statistical gap in scores between persons of color and whites is enormous. It is not. Depending on the test the difference varies but hovers in the range of 10 %.¹⁴ This difference in average scores has persisted over time, regardless of the type of test, whether it is a 'IQ' test norm-referenced or proficiency test, regardless of a test's publisher, or educational level of the test-taker, be it kindergarten or graduate school.¹⁵

¹⁴ Most of the data are from court records in Civil Rights challenges to standardized tests. See Walter Haney, George Madaus, and Amelia Keritzer *Charms Talismanic: Testing Teachers for the Improvement of American Education*, Review of Educational Research 14, 1987.

¹⁵ See Robert Linn 'Assessments and Accountability,' *Educational Researcher* 29:2 2000. He cites data from the Florida high school competency test, given annually since 1977 to illustrate a common pattern. When first a test is introduced, scores rise markedly for several years for whites and persons of color, level off, and over time decline slightly. However, the gap in test performance between the races remains virtually constant over time.

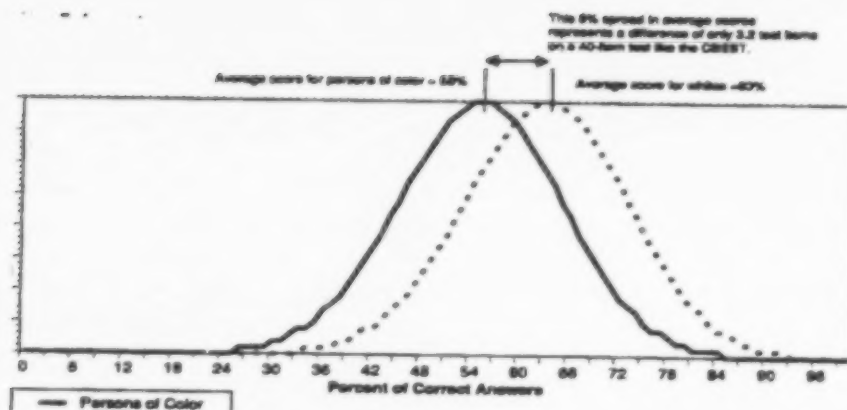


Figure 1 illustrates graphically an 8% difference in California's CBEST, a standardized test of basic literacy required of principals, teachers, and virtually any adult who works professionally with children in school. It is first important to note that the distributions of scores are highly overlapping. In practical terms, the difference gap amounts to a mere handful of test items. In Figure 1, the gap is an average of 3.2 multiple-choice test item on a fifty item multiple-choice test.¹⁶ (The number will vary according to the number of items on the test.) From an educational point of view, such differences have little if any significance. Because of the way the tests are normed and cut scores set, however, minor differences in the number of correct answers on a multiple-choice test create grossly inflated failure rates for persons of color. On CBEST, for example, African -American test-takers are 3.5 times more likely to fail the test than whites, Latino/ Hispanics more than twice as likely, and Asian Americans more than 1.5 times as likely to fail than whites.¹⁷

Figure 2

Number and Percentages of First-time Failures on CBEST 1985-95

	<i>Number eliminated</i>	<i>percent failing</i>
African Americans	11,200	63.0
Latinos	15,600	50.6
Asians	23,800	47.0
Whites	125,900	19.7

¹⁶ In CBEST, 10 of the 50 items on the math and language section are not scored. They are used in creating items for future version of CBEST. Eight percent of 40 items equals 3.2 items.

¹⁷ On some tests, particularly in mathematics and engineering, some Asian populations outperform whites.

Numerous researchers have carefully documented the highly disproportionate adverse impact on students of color of standardized achievement testing.¹⁸ An argument might be made that these differences in test scores, while small, nevertheless represent real differences in performance, and that tests, though imperfect, eliminate those most likely to perform poorly at school or on the job. Steele's study suggests the opposite - that the more talented students are at greater risk of failure. There is no evidence to support the claim that standardized tests are valid and credible measures of academic achievement or intellectual capacity. Further, there is no demonstrable connection between observed academic performance and standardized test scores. Test scores do not predict future success in school, the university, or in the workplace. Some standardized tests, the SAT for example, do correlate statistically to future grades. But this correlation is short lived.¹⁹ What standardized achievement tests appear to predict best are parents' wealth and scores on other similarly constructed tests. As reported by Peter Sacks, socio-economic class accounts for approximately 50% of the variance in SAT test scores. He estimates that for every additional \$10,000 in family income, a person on average gains 30 points on the SAT.²⁰

Among the more commonly heard explanations for the gap in standardized test scores is that the tests themselves are culturally and racially biased. What this has usually been taken to mean is that the bias is lodged in the content or language of individual test items. In the early years of mental measurement, the racism of the test items was often blatant. In more recent years, major test publishers have made efforts to review and eliminate items with overt cultural and racial bias. Though item bias remains, it is implausible to conclude that all the publishers in all their tests knowingly or unknowingly managed to create tests with an almost identical ratio of biased to unbiased items. The fact that scores on all commercially produced tests show the same eight to ten percent gap suggests that the gap cannot be fully explained by racial or cultural bias lodged in individual test items. Rather, the bias is *systemic* and *structural* - that is, built into in the basic assumptions and technology of standardized testing in the way the tests are constructed and administered, the way results are reported, and in the organizational structure and administrative rules of the accountability system itself.

There is perhaps no clearer illustration of how the differences among the races are greatly exaggerated and distorted than the numerical scales used to report results. There is, as I have noted, about a 10 percent difference in scores between white and

¹⁸ These include: Linda McNeil *Contradictions of School Reform; The Educational Costs of Standardized Testing* New York: Routledge, 2000; George F Madaus, 'A Technological and Historical Consideration of Equity Issues Associated with Proposals to Change the Nation's Testing Policy,' *Harvard Educational Review*, 64:1, 1994; Diana C Pullin, 'Learning to Work: The Impact of Curriculum and Assessment Standards on Educational Opportunity,' *Harvard Educational Review*, 64:1, 1994.

¹⁹ Some achievement tests, the college entrance SAT for example, predict academic grades at the next level, but only in the very short run.

²⁰ See Peter Sacks, *Standardized Minds*. Cambridge, MA: Perseus Books 2000.

²¹ The concluding paragraph, slightly revised, is taken from Deborah Meier, 'Educating a Democracy: Standards and the future of public education', *Boston Review*, Dec 1999/ Jan 2000.

nonwhite students. On a 100 point scale, this 10 percent difference constitutes a gap of ten points. However, California's Academic Performance Index or API (which is based entirely on students' scores on the Stanford 9 achievement Test) creates a 200 to 1000 point scale, and a 10 percent difference in scores morphs into a formidable 100 points. The SAT, the most commonly used test for college admissions, and also frequently used (inappropriately) to rank states' academic performance, creates a 400-1600 scale. On this scale, a ten percent difference is transformed to a 120 point chasm.

A major goal of social reformers of the 20th century was the elimination of legalized segregation. We still live in a society that is separate and unequal. To achieve social and economic justice, the goal for the 21st century must become the elimination of institutionalized racism in all sectors of social, economic, cultural, and political life - in business, housing, employment law enforcement, the courts, health-care institutions, and, of course, schools. What makes institutionalized racism so pernicious and difficult to eradicate is that racist practices are often invisible because they are accepted as standard operating procedures within our institutions.

Standardized tests are a particularly invidious form of institutionalized racism because they lend the cloak of science to policies that have denied, and are continuing to deny, persons of color equal access to educational and job opportunities. An educational accountability system based on standardized testing though predicated on 'standardized' measurements which are purportedly neutral, objective, and color-blind, perpetuates and strengthens institutionalized racism.

3 Tests and Multiculturalism

Tests are the single most important influence on the content of the school's curriculum and how it is taught. All tests, including those composed by teachers at the school level, confer upon students' attitudes, ideas, and images of what matters, and just as important, what does not. But the shift in power over the assessment process from the school and local community to state government represents a momentous and qualitative change. The power of the state over rewards and sanctions imposes singular answers to the questions of what schools are for, what constitutes genuine knowledge and learning, and what the next generation should and should not learn at school.

Whatever does not contribute directly to short-term gains in test scores is marginalized -- critical thinking, interdisciplinary studies, music, the arts, physical education, and forms of multicultural curriculum and bilingual education that are not add-ons, but integral to the entire curriculum. Even if tests incorporate some multicultural, and multiracial content, state control undermines local efforts by parents, teachers, principals and elected officials to revitalize their local schools, rethink curriculum and pedagogy, and respond in ways that cultivate the major assets of a multicultural society -- its racial and cultural diversity and a heterogeneity of perspectives on knowledge, culture, and learning.

Finally, among the most serious negative educational consequence of high stakes state-mandated tests is that teachers and administrators in low scoring schools are under such extraordinary pressure to raise test scores that those most likely to be first in line for a narrow and culturally truncated curriculum, and the recipients of shrinking educational opportunities are the children of the poor, immigrants, and people of color.

While Presidents Clinton, Bush, and other defenders of the excellence via testing policies are never heard proclaiming that one of the chief purposes of government mandated testing and indexing policies is to employ government power to unify the culture, it is clear that from the beginning that this has been a chief corollary goal of the architects of these policies. The seminal 1992 report, *Raising Standards for American Education* that launched Goals 2000 argued that testing tied to national standards would 'bind together a wide variety of groups into one nation,' providing '*shared values and knowledge*' which will serve 'as a powerful force for national unity.' Lauren Resnick, an academic advocate for 'smart tests', a former president of the American Educational Research Association and one of the originators of the New Standards Project, argued, 'Without performance standards, the meaning of content standards is *subject to interpretation*, which if allowed to vary would undermine efforts to set high standards for the majority of American students' (italics added). Nicholas Tate, chief executive of the British government's 'Qualifications and Curriculum Authority' is more forthright. He said in an interview, 'Today, we face the widespread belief that there are no underlying shared values in our society, that people are no longer willing to go along with what the school says. That is why we are beginning to make explicit, what has hitherto been implicit.'

It is no coincidence that this concerted effort by governments to gain near monopoly control over the curriculum arrives at the time that social movements have appeared and are challenging the cultural dominance of western Anglo-European traditions in the curriculum. The multicultural, bilingual movements are expressions of the will of men and women of differing races and ethnicities, an assertion of their rights, which includes the right to reclaim cultural power, and to forge their own cultural and social identities. But some see the diversity and heterogeneity of multicultural and bilingual movements as a threat to national unity, fostering balkanization of the nation, and the erosion of culture and academic standards. Ironically, the Goals 2000 plan of government imposed common curriculum tied to mandated testing as a way to foster social stability and promote national unity achieves the opposite. In practice, it exacerbates inequalities and provokes racial, cultural, and social strife.

As the locus of control of the assessment system (with its sanctions and rewards) shifts power upward, not surprisingly, new arenas for political culture warfare are opened at the national state levels, within the bureaucratic apparatus of the executive, legislative, and judicial branches of government. Decisions are being made increasingly by politically appointed 'blue ribbon' commissions, panels of experts and consultants many degrees removed and insulated from local community concerns, and remote from children, teachers, classrooms, and schools. What best serves interests of this child, this classroom, and this community is lost as control of the curricular and pedagogical decisions shifts to the upper reaches of government, to state capitals and to Washington where the 'stakeholders' are those who are well organized, have the financial resources and power required to compete in the political arena.

As the decisions are made farther and farther away from classrooms and schools, differences over controversial moral, cultural, and educational questions are magnified and intensify, and (as in California) likely to become entangled in acrimonious, racially charged and xenophobic electoral politics.

The kinds of learning required of citizens in the modern world cannot be achieved by standardized and centrally imposed systems of learning. Human learning to be effective and long-lasting requires the engagement of the learner on his or her own behalf, and rests on the relationships that develop between schools and their local communities, and between teachers and their students. Powerless school communities and teachers cannot produce powerful, engaged citizens committed to social and racial justice, and the public good.¹

¹ The concluding paragraph, slightly revised, is taken from Deborah Meier, 'Educating a Democracy: Standards and the future of public education', *Boston Review*, Dec 1999/ Jan 2000.

4 Standardized Achievement Tests and Race

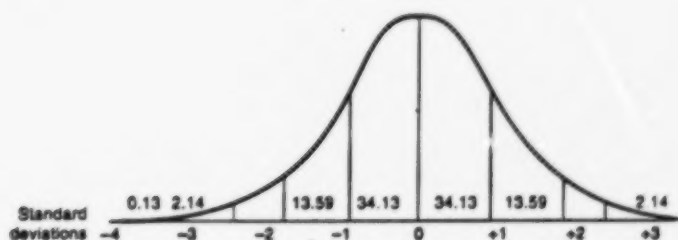
How credible and dependable are mass administered standardized tests as measures of academic achievement? All of us brought up and schooled in America are familiar with standardized tests. We sit in a classroom or auditorium well spaced from fellow test-takers, given a time limit and booklet of test 'items'—passages of text, math problems, tables, diagrams or charts. From four or five possible responses, we choose the one correct or best answer, and darken the appropriate bubble on the answer sheet. Some standardized tests include 'open-response' items, which are scored according to standardized procedures.

Broadly speaking there are two types of mass-administered, standardized educational tests used in current accountability systems. The first are 'norm referenced' tests (sometimes referred to as 'standardized' tests). These tests do just that; they create norms, percentiles or grade equivalency scores which indicate an individual's or group's standing on the bell curve relative to all others taking the same test.

The second type are 'proficiency' or 'basic skills' (sometimes called criterion-referenced) tests which employ either a pass/fail or proficiency level 'cut' score.

From the point of view of the test-taker, the 'norm-based' and most 'criterion referenced' are indistinguishable. They look alike, contain very similar test items, and are generally given under the same time constraints and conditions.² Both use standardized scoring and reporting procedures and rely on 'normal' or bell curve statistical models. The major difference is in the way each creates failure. A norm-referenced test is deliberately constructed and pre-tested to yield scores distributed so as to approximate the so-called 'normal' curve. The curve defines fixed percentages of high, medium and low-test performers. (See Figure 3) By definition, 50% of the test takers fall above, 50% below average. About 70 percent fall somewhere in the middle, with about 15% high above and 15% far below the mid point.

Figure 3



The technology of the norm-referenced test accepts as given the highly dubious assumption, that (if the sample is random and large enough) all human qualities, traits, capacities, achievements, etc., if properly measured will approximate the bell-shaped 'normal' curve. Failure is created by the bell curve. With a proficiency or 'basic skills' tests, failure is created by a particular cut or proficiency score selected by a group of human beings, elected or appointed government officials and/or a panel of experts chosen by these officials or by the testing company under contract to a state agency.

The most fundamental problem with both types of standardized academic achievement tests is that there is little evidence to support the contention that they measure what they purport to measure –academic achievement, or proficiency. This does not mean that academic achievement and high standards are not vital. Rather it is that the tests have very little relationship to actual academic performance of any kind. For some standardized tests there is correlation to grades, at least in the short term. But, for virtually all standardized reading and writing tests, there is no demonstrable connection between a person's performance on a standardized reading test and a person's reading abilities in the real world –in everyday life situations, at school, work and elsewhere where one might be called on to read. *What this means is that contrary to common sense, Student A's score at the 45th percentile and Student B's at 95th percentile on the Stat-9 reading test (or any other norm-referenced tests) says nothing whatsoever about the actual or relative reading performance of students A or B.* The standardized test informs us only how every test-taker's score on the test compares to everyone else taking the same test.

A score on basic skills' or 'proficiency' test tells us only how far above or below the established cut-off a student's score falls. Cut scores on academic proficiency test are *not* based on actual or observed level of competence or proficiency. There have been numerous studies that have explored the relationship between test performance and actual performance, and researchers have repeatedly come up with the same conclusion: no (or almost no) connections. Neither do the tests meet the criterion of 'predictive validity'. Norm-referenced and proficiency tests (except for grades in the short term) do not predict future success in school, the university, or in the workplace.³ What the tests predict best is a person's score on similarly constructed tests and parent wealth.

The failure to ground cut scores in performance applies to California's Academic Performance Index (API). The score of 800 established as the mark of excellence for all schools to aspire to reach and exceed is a wholly statistical construct. It is not based on any direct observational evidence or documentation. It is extraordinary and also sadly ironic that the cut score now driving state educational policy for achieving educational excellence is not grounded in any way on educational excellence and high academic achievement *as they are manifest in the real world of teaching and schools.*

2 Some states afford flexibility in time allotted for completing the mandated test. In some cases allowances are made for students with documented learning disabilities, or whose native or first language is not English.

3 Some achievement tests, the college entrance SAT, for example, predicts academic grades at the next level, but only in the very short run.

The seriousness of the problem of the failure to ground standardized tests in actual performance is now widely acknowledged. Clinton's Secretary of Education, Richard Riley, a longtime Goals 2000 testing enthusiast, in his final 'State of American Education' address urged the states to stay the course, but also cautioned state officials about the dangers of relying on a single test for making high stakes decisions. The Secretary's caution is a response to internal and external pressures—including the US Department of Education own Office of Civil Rights (OCR) which has issued guidelines which assert that the use of test scores as the single factor to determine retention, graduation, and college admission is improper and possibly a violation of Civil Rights law.

The OCR guidelines' are grounded in two recent studies conducted by the National Research Council of the National Academy of Sciences.⁴ These two studies are but the tip of the iceberg. There is a vast literature in mainstream psychological and educational measurement research that raises fundamental question about the meaning and usefulness of norm-referenced and conventional standardized proficiency tests.

For the first time in its 50 year history, the 1999 revision of *Standards for Educational and Psychological Tests* produced jointly by the American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) indicates that the validity of educational tests cannot be established without reference to how they contribute to the improvement of student learning and consideration of the negative consequences of test use. Further, the standards assert that no "decision or characterization" of students which has major impact on their future should be made on the basis of a single test score, and caution against the use and interpretation of tests for students with learning disabilities and with limited English language proficiency.

State mandated testing and indexing policies that distribute rewards and sanctions based on test results are common and are increasing. This in the face of almost universal agreement among independent experts on the technical limitations of standardized achievement tests and that their use as a high stakes measure of educational achievements or capacities is destructive, misleading and inappropriate. Government regulations and mandates linking tests to high stakes decisions proliferate, at the same time that the standing of standardized tests' as trustworthy instruments of modern social science has never been lower. Why are indexing policies that strengthen state control so readily endorsed and supported by politicians, corporate leaders, and national teachers unions? This question is addressed in the final essay that explores possibilities for fundamental changes in direction in assessment policy.

Because tests have long been used to determine merit and access to top academic tracks, special programs and high status schools and universities, they have been challenged over the years politically and in numerous suits alleging violations of US

⁴ *High Stakes Testing for Tracking, Promotion and Graduation*, 1998, and *Myths and Tradeoff: The Role of Tests in Undergraduate Admissions*, National Research Council. 1999.

civil rights law. Much of the heated controversy over affirmative action also rests on the continuing use of standardized testing to define merit.

5 Effects of increased government control of assessment

An independently funded set of studies conducted by the *Centre for Assessment Studies* at the University of Bristol, UK sought to map empirically the consequences of the Education Reform Act of 1988, initiated by the Conservative government led by Prime Minister Margaret Thatcher.⁵ This law created a nationwide school indexing system for England and Wales, and shifted control of curriculum from individual schools, school councils and local Educational Authorities (school districts) to a central government authority called OFSTED (Office of Standards in Education). A team of researchers over a period of eight years (1989-1997) studied a national sample of primary schools employing a wide range of systematic social scientific qualitative and quantitative methods. The study produced dozens of articles and four major books, the most recent published in late 2000.⁶

The researchers document the grand mismatch between policy intentions and the outcomes. Rather than erasing educational inequalities and raising the level of academic accomplishment as promised, the state-mandated assessment process served to obstruct learning, perpetuate and increase disparities. Tests, even 'good' tests served to distort and disrupt learning, in particular for bilingual students whose first language was not English. Documented also was a dramatic narrowing of the curriculum and restriction of the range of learning opportunities, increased devaluation of teacher knowledge, decline in teacher and headteacher (principal) morale. There were also increases in pupil anxiety and dysfunctional changes in school structure and governance. This included various forms of resistance by teachers and administrators to somehow hijack the rules and circumvent the system. They engaged in behaviors some might call 'cheating,' and others principled defiance of government regulations that denied opportunity and stunted student learning. Other studies have documented that headteachers once highly independent and insulated from the twists and turns of national electoral politics became highly vulnerable. As a consequence, it has become increasingly difficult to recruit and retain creative and talented teachers and administrators in schools located at the bottom of the school rankings --which, in England, as in the US, serve children of the poor, immigrants, persons of color, a majority of whom live in the nation's most distressed urban areas.

5 The 1988 Education Reform Act proposed and shaped while Margret Thatcher was Prime Minister became law under her successor, John Major.

6 Pollard, Broadfoot, Croll, Osborn and Abbott, *Changing English Primary Schools*, Cassell, 1994; Croll (ed.) with Abbott, Black, Broadfoot, Osborn and Pollard) *Teachers, Pupils and Primary Schooling*, Cassell, 1996; Osborn, McNess and Broadfoot, with Pollard and Triggs *Policy, Practice and Teacher Experience* (Continuum 2000; Pollard and Triggs, with Broadfoot, McNess and Osborn *Policy, Practice and Pupil Experience* Continuum, 2000.

Though there are structural differences between the English and US systems, this set of studies is significant and relevant because California's API is modeled on Britain's 1988 School Reform Act. In 1998, in the waning months of his administration, the press reported that the outgoing governor Wilson had tea with former Prime Minister, now Lady Thatcher. He reported that his proposal to establish the API was based on his great admiration for the centralized system of school accountability she had installed. The study is also important because it leaves little doubt that the central issue is control. The negative effects of centralized curriculum control and indexing are evident regardless of the quality of the assessment instruments used. By US standards, many of the British assessment tools would be considered 'smart' tests. Thus, even if the flaws built into standardized tests such as the Stat-9 could be remedied, or if the tests were tomorrow replaced with a new generation of 'authentic' tests, there is little doubt that the negative effects of centralized state control of the assessment system carefully documented by the British researchers would remain in place. The British findings are also particularly instructive to those reformers who focus almost entirely on eliminating or correcting the deficiencies of standardized testing, and ignore both the race question, and the fundamental issue of power—who controls the assessment system controls the curriculum.

There is no comparable comprehensive, longitudinal study of the impact of a policy of government curriculum mandates in the US. There is, however, a large body of research on the use of standardized tests in making high stakes decisions. Numerous studies have been conducted by researchers from some of the nation's leading educational research universities, and independent R&D centers devoted to evaluation, testing and assessment. In a the *Educational Researcher*, the lead journal of the American Research Association, Robert Linn who is among the nations respected experts on educational testing, commented on fifty years of research on the use of tests and assessment in accountability systems. He concluded that 'common standards and testing encourages a narrowing of educational experiences for most students, dooms many to failure, and limit the development of many worthy talents.' This, he adds, 'should not to be misinterpreted to mean that one should not have high standards for all students...[H]aving high standards is not the same as having common standards for all.' Professor Linn, concludes with a damning commentary on the state of the science of educational measurement.

As someone who has spent his entire career doing research, writing and thinking about educational testing and assessment issues, I would like to conclude by citing a compelling case showing that the major uses of tests for student and school accountability during the past 50 years have improved education and student learning in dramatic ways. Unfortunately, that is not my conclusion. Instead I am led to conclude that in most cases the instruments and technology have not been up to the demands placed on them by high stakes accountability... Assessment systems that are useful monitors lose much of their dependability and credibility . . . when high stakes are attached to them. *The unintended negative effects of high stakes accountability... outweigh the intended positive effects.* [Italics added]⁷

7 Robert Linn, *op cit.*

Recent Studies in Texas and Massachusetts confirm Professor Linn's conclusions. These states have adopted a policy of linking test results to school promotion and high school graduation. While George W. Bush and his Secretary of Education, former Houston school superintendent Rod Page, point to higher test scores, the actual numerical gains in test scores were inconsequential, and a subsequent Rand study cast considerable doubt on the claimed gains.⁸ Boston College researcher Walt Haney showed that a sharp increase in the numbers of students retained in grade nine inflated Houston's high school passing rates, and that the numbers of dropouts in Houston, already among the highest in the nation, soared, most markedly for African Americans and Mexican Americans.⁹ A study by Rice University Professor Linda McNeil documents the multiple ways Texas policy pushed students out of school, increased educational disparities, and degraded the curriculum as a consequence of enormous institutional pressures to prep students for the state mandated standardized test.¹⁰

Massachusetts is in the process of imposing MCAS, a statewide achievement test, as a condition for receiving a high school diploma. A coalition of educational and community activists, civil rights and child advocates managed to delay full implementation. An independent study concluded based on a trial run of 10th graders, that the failure rate in schools and districts that predominantly serve the poor and people of color could reach and exceed 70%.¹¹ In Michigan a similarly constituted coalition succeeded in defeating the effort by the state to impose a high stakes testing policy.

It is impossible to sustain the argument that policies that have been shown to degrade curriculum and pedagogy, increase drop-outs and exacerbate inequalities, and that have no known educational benefits, will improve the level of education of the nation's youth or enhance their chances of competing in new global economy. It is also a debasement of the social and behavioral sciences, when the observations and judgments of those with direct first-hand experience, parents, teachers, principals, counselors, teaching and learning specialists are overridden, and dismissed as scientifically unfit and subjective, while standardized tests are valorized as the one and only scientifically valid measure of academic performance.

8 At best the gains are mixed. California reports 4-5 percentile points on the Stanford 9. Texas reports as much as 11 percentile points gain on its own test (TAAS). A recent Rand report *'Improving Student Achievement: What State NAEP Scores Tell us?'* (available at <http://www.rand.org>) shows gains of three percentile points or less. On the other hand, the Nation's Report Card compiled by National Center for Educational study indicates a small but steady decline in NAEP reading scores of high school students. (available at <http://www.nces.ed.gov>)

9 Haney, Walter, "The Myth of the Texas Miracle in Education," *Educational Policy Analysts Archives*, Vol. 8, No 41, <http://epaa.asu.edu/epaa/v8n41/>

10 Linda McNeil, *Contradictions of School Reform: Educational Costs of Standardized Testing*. New York: Routledge, 2000,

11 FairTest / CARE (Coalition for Authentic Reform in Education). (2000). *MCAS: Making the Massachusetts Dropout Crisis Worse*. MCAS Alert, Sept:

<http://www.fairtest.org/care/MCAS%20Alert%/>

6 The Alternatives and an Agenda for Change

The cornerstone of the Goals 2000 standards movement—raising educational standards by central government mandates that tie test scores to rewards and sanctions—is self defeating. The centralization of authority and the proliferation of standardized testing which has become pervasive in the past decade have shown no evidence of positive results. Indeed to the extent to which these policies have been implemented, there is substantial evidence that they have more often than not served as an obstruction to the pursuit of educational excellence and equity. To repeat the words of Robert Linn, one of the Nation's leading test experts, 'the evidence indicates that the unintended negatives of high stakes accountability systems probably outweigh the intended positive effects.'

This conclusion, however, is not to be mistaken as a rejection of the importance of raising standards and public accountability, nor as rejection of the need for national and state and local governments to use executive, legislative, and judicial power to protect student, parent, and community rights—and to take a strong affirmative role in the pursuit of equality and maintenance of high educational standards for all. Also, the fact that current national and state school reform policies are almost totally reliant on an arcane and deeply flawed test technology does not in any way diminish the need for accountability nor for effective and appropriate forms of testing and assessment.

Disputes over educational reform and accountability are often, sharply polarized, cast in terms of top-down vs. bottom-up, two apparently contradictory perspectives and sets of remedies for reforming and assessing the nation's schools. Government mandated testing linked to a uniform system of rewards and sanctions is the defining example of the former. The bottom-up view stresses local school and community-based initiatives, rooted in face-to-face encounters among teachers, principals, and parents in collaboration with community and local officials.

An assessment system in fact must serve both, 'top-down' and 'bottom-up' functions. On the one hand, it must provide dependable information to school authorities, advisory and governing boards, state legislators, local officials, etc., so they may be better informed to make policy decisions about the distribution of public resources. A systematic assessment process is key for holding districts, school officials and teachers accountable for the quality of their performance. On the other hand, the system of assessment must also provide information that serves the educational needs and interests of each individual child, strengthens local school and community level reform initiatives aimed at improving teaching and learning, and cultivates the integration of diverse cultural historical perspectives and language traditions into the school's curriculum and pedagogy. To serve the nation, and serve children of diverse cultural, racial, ethnic, language, and religious traditions, there must be an appropriate balance of power between central government authority and local school and community control. And how the power over rewards and sanctions is distributed is key.

A more, balanced, inclusive, effective and democratic set of national and state educational assessment policies is workable, possible, and not beyond reach. A shift in power would temper and moderate the already highly disproportionate power held by federal and state government authorities. A shift would restore balance by giving significantly more voice and greater responsibility not to state governments but to the 'grassroots'—to individual schools, teachers, parents, and local communities.

What follows from a shift away from the center is that many of the differences rooted in fundamental moral religious, and cultural beliefs, including philosophical and ideological differences related to curriculum, pedagogy and learning, would be resolved face-to-face, by locally constituted groups in an open consultative process, rather than by panels of experts, executive or legislative commissions appointed by state and federal officeholders and far removed from local communities, schools, children, and teachers. The resolution of the basic dilemmas of teaching, learning, and curriculum would be distanced from divisive, xenophobic, electoral politics.

Obstructions to change

Redressing the power imbalance is possible but by no means assured. The political and institutional support for current centralized policies is strong, and the resistance to rethinking and reformulating of national and state assessment policy is considerable.

Though the Goals 2000 standards movement shows no promise of producing excellence and equity as its proponents promised, it has wide support by the public and among politicians from the President on down. Why, despite the intense criticism and in the face of increasing resistance by students, teachers, and community activists,¹² does the pro testing/standards perspective continue to have such a strong hold on popular opinion, and remain dominant?

The cultural / psychological barriers to rethinking assessment policy and practices are formidable. As a culture, Americans believe in tests, standardized tests in particular. Tests used for making high stakes decisions have a deep psychological hold on us because we are part of, and surrounded by, a culture where the need to assign numbers to performance and to compare and rank order individuals and institutions is seen as self-evident. In a world where rank and test scores matter, we also assume that test scores will tell us whether our children are prepared to compete in the hard, cold world. We also want our children's local school, and school district to be among the best. In addition, to a lesser or greater degree, most of us schooled in this society have come to accept standardized tests as a measure of our self worth, particularly with respect to our estimate of our and our children's intellectual and academic capacities and abilities. For many Americans, an educational system without or with a very greatly diminished place for standardized tests is inconceivable.

¹² Movements of teachers, community and youth activists in opposition to tests are gathering force across the nation including California, Massachusetts, New York and Illinois. There have also been boycotts by students in Chicago, and Boston.

Any remedy that would disperse power downward to schools and communities will be greeted with skepticism. Though Americans celebrate democracy and democratic values, when confronted by difficult problems or a crisis, as a nation and culture, we are inclined not to more democracy but less. The political and popular culture more readily endorses solutions that promise immediate measurable results and that rely on hierarchical power relations backed up by a universally applied system of rewards and punishments.

There are also political and economic obstructions to rethinking and reforming assessment. A whole generation of mainstream politicians, presidents, senators and governors (e.g. Bill Clinton, Lamar Alexander, Richard Riley, George Bush 1&2, virtually all the 'new' Democrats, Evan Bayh, Al Gore, Joe Lieberman, many state education officers, state politicians, corporate, the national AFT (American Federation of Teachers) leadership and remain fixed on a "standards" movement predicated on national testing and raising test scores. Despite its failure, the policy persists, in part because they have no other solutions to offer. Also, those now and formerly in office who were responsible for conceiving of and instituting these policies are not likely to concede that the Goals 2000 high stakes testing scheme has had disastrous consequences and is a failure.

Furthermore, there are vested interests. Testing is big business. The major test producers, distributors and service providers are divisions of large publishing firms, which in turn are divisions of major corporate media conglomerates. US Bancorp estimated that George Bush's educational testing proposals (which as I write, are marooned in Conference Committee of the U.S. Senate and House) would more than triple the current \$300,000 annual expenditures of direct cost of tests. The National Association of State boards of Education estimated that the new testing mandates alone would cost as much as 7 billion.¹³ Costs include loss of teaching days, expenditures for salaries, physical and equipment and facilities to support an army of experts, consultants, state and district-level bureaucrats, legal advisors, testing experts, whose livelihoods and careers depends in whole or in part upon administering, scoring tests, reporting, storing analyzing, classifying test results, and insuring compliance with federal and state laws and test regulations. Almost ten years ago, researchers George Medaus and Walter Haney at the Center for Evaluation and Policy Research at Boston College estimated these costs at a 20 billion annually.¹⁴

Finally, among the more formidable obstructions to change in assessment policy is a belief, widely shared in this society, that whatever the deficiencies of high stakes standardized testing policies, there are no other practical ways to raise standards, assess educational progress, sort students and evaluate teachers. In spite of the enormous direct and indirect costs many still assume there are no alternatives or at least no economically feasible alternatives to existing policy. They may concede the imperfections of tests, but insist they are the best we now have. Without centralized

13 M. Fletcher and N. Irwin 'Public Reform, Private Windfall. Washinton Post, pp A-09 Aug 16, 2001

14 Walter Haney, George Madaus, and Robert Lyons, *The Fractured Marketplace for Standardized Testing* Boston: Kluwer, 1993.

testing, the system of elementary and secondary education would be unaccountable, aimless, and without high standards.

Policy alternatives

The educational policy issues surrounding testing, assessment, and public accountability are immensely complex. I do not offer here a blueprint for reform of the system of accountability that will tomorrow overturn and repair the damage created and fostered by Goals 2000 policies. I do, however propose several guiding principles for assessment reform, and three fundamental issues that must be addressed in formulating and pursuing alternatives. I closes with recommendations for shifting the balance of power in order to create a fair and effective accountability system.

Principles:

A fair and effective accountability system will:

1. help to achieve and maintain high educational standards, but will not seek to standardize the curriculum, the learning process, nor attempt to impose a singular view of knowledge, language and culture;
2. contribute to the education of the nation's children to the full range of their talents and capacities;
3. serve to assure more equitable distribution of resources and equality of access to educational and job opportunities;
4. serve to encourage and reward initiative and meritorious performance --schools, teachers and students;
5. contribute to erasing institutional racism.

Three key issues to address

Institutional Racism

Virtually absent in discussions of educational reform by mainstream press political leaders is mention of the pervasiveness of institutional racism, and of the enormous inequalities in human and material resources between schools that that serve white children and those that serve predominantly children of the poor, immigrants and of color.

It is of vital importance that the accountability system specifically address the legacy of white supremacy and institutionalized racism legitimated by standardized testing, a legacy that lives on in the present. Many who support the standards and testing movement undoubtedly have good intentions, but whatever the intentions, the policies they helped to enact serve to sustain and strengthen institutional racism.

The institutional racism is evident not only in the outcomes --the disproportionate number of persons of color who fail or are eliminated based on standardized test scores-- but is built into the assessment technology itself. That is to say, built into the assumptions, instruments and the methodology of standardized testing, a primitive technology that dates to the early years of the twentieth century. Institutional racism is manifest in who has the political power and who is under or unrepresented when fundamental decisions are made about allocation of resources, choice of curriculum

content and teaching methods, eligibility for programs, grade advancement, and the award of educational credentials. Key is power to set the rules, name the 'stakeholders,' allocate the resources, and make the final decisions.

An accountability system to be fair and effective must make affirmative efforts to counter the institutional racism currently built into the technology of the instruments of assessment. Procedural and structural protections against institutionalized racism depend on proportionate distribution of decision-making power with a significant degree of political and cultural control vested at the school and community levels.

Technology.

Contrary to a widely held belief, there is no shortage of systematic evaluation instruments for assessing teaching and school learning and for gauging the quality of 'academic' and other forms of school learning.¹⁵ Some of the 'alternatives' are highly developed and have been shown to provide teachers, parents, and local officials with useful information for enhancing student learning and/or making local and internal school policy decisions. Some of these approaches are more cost efficient than conventional standardized tests because the time spent on assessment is not time lost but integral and additive to the teaching and learning process. It is also important to note that the use and interpretation of these instruments is dependent on the social context and particular situation. Thus, none are suitable for producing a single numerical scale that serves as a universal measure of the educational productivity for all situations, schools, teachers, and students.

Note that the technology of multiple choice standardized testing as we now know it, was developed in the first two decades of the 20th century at a time when mechanical hole punch and manual sorting with pins was state of the art information processing technology. The high-speed digital microprocessor and desktop computer technology developed over in last decade has transformed our technological capacity to collect, process, organize, and use very complex information. Other than the introduction in the 1930's of machines capable of reading the graphite pencil marks on answer sheets, and their replacement with digital scanners in recent years, the basic science and technology of standardized test taking and reporting is virtually unchanged since it was invented nearly a century ago. By today's standards, the technology represented in the *Stanford t-9* and all similar tests is dated, limited, and highly static.

It is not likely that the innovations in testing and assessment technology will originate from the testing industry, which is heavily invested in multiple-choice technology and ill- equipped for dealing with the new cutting edge information technologies. They have little to gain and much to lose from a accountability system that does not rely on

15 Deborah Meier *Will Standards Save Public Education* Boston: Beacon Press, 2000; Linda Darling-Hammond, J. Aneess and B. Falk, *Authentic Assessment in Action*, New York: Teachers College Press, 1995; Grant Wiggins, *Educative Assessment: Designing Assessment to Inform and Improve Student Performance*, San Francisco: Jossey-Bass, 1998; Patrick Griffin, P. Smith and L. Burrill *The American Literacy Profile Scales: A Framework for authentic Assessment* Portsmouth NH: Heinemann, 1995; Monty Neil et al. *Implementing Performance Assessments* Cambridge MA: FairTest, 1996.

centrally administered and scored standardized multiple-choice tests. Though no technology can replace human judgment, the newer digital information technologies do have unexplored potential for fostering responsive, systematic, and locally based assessments —assessments that also teach. To avoid commercialization of the educational process and undue influence of large corporate interests, pursuing these paths will require public investments that stimulate school and community-based collaborative research and assessment development.

In the near future a variety of educational tests will continue to be used in diagnosing student needs and assessing educational achievements. There are a number of steps that ought to be taken immediately by governments to protect children, communities and the public at large from discriminatory tests, and insuring that the tests used meet the dual standard of enhancing learning and advancing equality of educational opportunity. Forms these protections might take is briefly discussed in the concluding section.

Political Power and the Assessment System

How should political power distributed within the accountability system, that is who should write the rules, possess the power to distribute rewards and sanctions, determine who are to be the experts and 'stakeholders' who in the end make the fateful educational decisions? The instruments employed by the system of accountability are of course critical, because they define what is valued. But an accountability system includes a distinct organizational structure, a set of procedures for controlling rewards and sanctions that represent a particular configuration and distribution of power. The configuration can be changed and balanced so as to give more weight and responsibility to schools, teachers, parents, and local communities, and less to panels of experts, legions of officials, major corporate and foundation executives, national and state commissions and agencies.

A Massachusetts group called the Coalition for Authentic Reform¹⁶ (CARE) has outlined a proposed statewide accountability system that aims to raise educational standards and the quality of learning and teaching in classrooms and schools. It consists of four integrated components.

1. *Local Assessments.* Each school would submit its accountability plan for review and approval to a regional board, established by the Massachusetts Department of Education. The plan would outline how the school will assess progress toward a broadly stated set of competencies.
2. *External Quality Reviews.* On a three to five year cycle schools would undertake a self study and an external auditors would review the self study visit the school, report on progress toward the dual goal of academic excellence and equitable and quality resources and learning opportunities are being provided to all students.
3. *Standardized tests.* These would be limited to literacy and numeracy and would not have high stakes decisions attached.
4. *Annual Reports.* Each school and school district would annually report to 'stakeholders' on a set of 'indicators' developed by the state. These would include

¹⁶ Full proposal and list of members of the coalition www.fairtest.org/care/accountability.html

but not be limited to academic performance and reported in terms of race, gender, low income status, special needs and limited English proficiency

The State of Nebraska in 1998 adopted policies that emphasize that the assessment of student academic performance is a local responsibility that should primarily serve to improve instruction and increase learning in the classroom. Further, the policy asserts that no single measure can achieve all purposes, and multiple measures are needed to provide complete information to teachers, parents and policymakers. The assessment system called School-based, Teacher-led Assessment and Reporting System (STARS) was set to begin in the 2000-01 school year. Under this plan,¹⁷ the Nebraska Department of Education invited proposals from teachers and local districts to develop their own operational plan. One of the proposals submitted and approved was from a coalition of representatives from the Nebraska Writing Project (the home of the National Writing Project), 'The School at the Center.' (Both are networks of teachers and university faculty members) and nine Nebraska school districts. The underlying premise of their plan is that teachers develop the assessments and become their own assessment experts. The Coalition promises to produce nine "locally appropriate, context-sensitive assessments" for mathematics and reading/writing. The Nebraska plan has its limits. Originally all districts could administer one of several state approved, commercially available achievements test batteries. In 2001, under pressure to conform to federal Goals 2000 'guidelines,' Nebraska now requires a single test. It still provides funds for local school and school district assessment initiatives, and places significant restrictions on the use of standardized tests in making high stakes decisions.

There are also other living examples of accountability systems where power is balanced between national and local interests and concerns. Scotland with population of just over five million (approximately the size of Maryland, Missouri, Wisconsin, Minnesota) governs its own primary and secondary schools independently of the British government in London. Traditionally education in Scotland is organized as a partnership between the central government, local government, and schools. For many years it has had a system of school inspection that resembles the school review process proposed by the Massachusetts CARE coalition. American style standardized tests play no significant role in the assessment process. Two government papers¹⁸ reasserted and strengthened the policy that it is the responsibility of the national executive authority 'to exert strategic leadership of the national system...by articulating after consultation the national priorities for education, yet leave to each school supported by its local authority [school district] the central responsibility for its own improvement and for raising standards.' Further, the paper affirms national policy that specifies a basic level of provision which specifies a minimum educational resources for all schools and students.

17 See the STARS Planning Guide at: <http://www.edneb.org/IPS/starsmnt.pdf>. Also see Chris Gallagher *A Seat at the Table: Teachers Reclaiming Assessment Through Rethinking Accountability* Phi Delta Kappan <http://www.kiva.net/~pdkintl/kappan/kgal0003.htm>

18 Scotland Executive Education Department, *School code paper and Priorities for schools* 2000.

There have also been several notable efforts in the US and in the U.K. to develop comprehensive accountability models that could serve to articulate local school-based systems into a national (or regional) assessment framework.¹⁹

An Agenda for change

1. Eliminate regulations that directly or indirectly link federal incentives to state adoption of centralized statewide testing of teachers and students.

2. Require an educational impact statement or 'report card' on the material costs, educational and social consequences of a test or assessment procedure by any government educational agency prior to implementation. Such a statement would report on effects on students, schools and community including academic performance, distribution of physical and human resources and learning opportunities, drop-out, etc. broken down by race, gender and socio-economic level.

3. Provide federal and state resources to schools that (with the support of parents and locally elected school officials) take central responsibility for school improvement and raising standards. Stimulate the development of partnerships among teachers, communities, parents, to develop 'locally appropriate, context-sensitive assessments.'

4. Strengthen and support efforts to set and enforce standards for tests and assessments that protects the public from inappropriate use of tests and assessments, violations of civil liberties and rights of students and teachers. Require disclosure of consequences and 'side effects.' Currently there are two sets of relevant standards: the *Standards for Psychological and Educational Tests* produced by three professional associations;²⁰ and *Principles and Indicators for Student Assessment Systems* developed by the National Forum on Assessment, a coalition of children's and national civil rights groups.²¹ Both are useful but largely symbolic since test developers and government agencies are under no legal obligation or political pressure to meet any standards or principles. The major test publishers often insist they meet professional standards but they operate under a shroud of secrecy that is sanctioned by the courts and difficult to penetrate. Remedying the situation will require legislation, or various forms of extra-governmental agreement that require that any tests used for high stakes decisions meet published professional standards, and that educational tests and assessment procedures be open to public scrutiny and independent review.²²

19 Ann Filer (ed.) *Assessment Social Practice and social Product* Falmer Press, London & New York, 2000; Harold Berlak, (ed.) *Toward a new science of educational testing & assessment*, Albany: SUNY Press, 1992; John Raven "A model of competence, motivation, and behavior, and a paradigm for assessment" in Berlak (ed.) *Toward a new science*; Tyrell Burgess and Elizabeth Adams *Outcomes of Education*, London: Macmillan Education, 1980.

20 The American Psychological Association, American Educational Research Association, and National Council on Measurement in Education).

21 Summary available at: www.fairtest.org

22 In 1998 the Ford Foundation funded an organization called National Board on Testing and Public Policy, located at the Center for Testing, Evaluation and Educational Policy at Boston College. One of its chief purposes of the Board would be to monitor quality by conducting independent expert audits of tests. One of the more serious difficulties with the proposed Board is that testing industry is elevated to be a major 'stakeholder.' This would almost

5. State legislatures should declare a moratorium on tests used high stakes testing in order to undertake a systematic review of the tests to determine whether they comply with professional standards and meet National Forum assessment principles.

6. Federal and/or state government could enact legislation to rein in the abusive uses of tests. Paul Wellstone (D-MN) in April, 2000, introduced to the Senate, and Representative Robert Scott (D-VA) to the House *The Fairness and Accuracy in Student Testing Act*,²³ that would have prohibited the use of standardized tests as the single determinant in making decisions about graduation, promotion, tracking or ability grouping of students. Test producers would have had the burden of proof to show that a test is valid and reliable for the purposes for which it was being used and that the tests fairly assessed what student were taught in school. Students would have been guaranteed multiple opportunities to demonstrate proficiency and appropriate accommodations were required for students with limited English proficiency or disabilities. The proposal went nowhere. Action on a state by state basis is more likely to be successful.

Changing assessment; changing schools

Goals 2000 policies has led the nation down a dangerous path by increments to a radical transfer of power, with increasing concentration in the hands of government, authorities, bureaucrats, experts, and Washington D.C. and state capital based 'stakeholders,' all distant from children, classrooms, and schools. This power imbalance is educationally, and as recent electoral politics of California illustrates, politically unwise, and potentially explosive. Mandated standardized tests because they have a disproportionately high adverse impact on communities of color, sustain and strengthen institutional racism. As testing programs that tie tests to school promotion, admission to 'gifted' programs, entitlement to high school, scholarships, diplomas, degrees, certification, etc., are implemented over the next several years, the adverse effects will intensify, and provoke racial and cultural conflict, and organized resistance.

Significant school reform is not possible without major reform of the current system of national and state educational assessment. Changes will not occur of their own accord. They will come about only in response to persistent pressure by coalitions and tactical alliances that cut across political, social class, racial, and ideological lines. These include coalitions of citizens, students, teachers, parents, child advocates, civil liberties and civil rights leaders. In 2001 there were numerous organized, protests, boycotts, and other forms of active resistance to high stakes standardized testing across the nation. Resistance is growing and becoming more militant as mandated tests tied to sanctions are put into place. We as a nation will continue differ profoundly on how schools ought to educate, what an educated person ought to know, and on how

certainly retard innovation by giving precedence to the companies who are heavily invested in an out-of-date technology. Also a problem is the absence of a significant presence on the governing board of practicing teachers, parents, and local community. See George Medaus and Cathy Horn, *Testing Technology: the need for oversight* in Ann Filer *op cit*.

23 <http://www.senate.gov/~wellstone/highstakes2.htm>

students learn best. In a democracy we cannot allow government, panels of experts remote from communities, classrooms and students to impose a singular view of curriculum and learning, and to decide our and our children's futures.

© 2001 Harold Berlak

Harold Berlak is an independent consultant and researcher, senior research fellow at the Applied Research Center in Oakland California, and a fellow at the Educational Policy Research Unit at Arizona State University. His most recent articles include "Race and the Achievement Gap" *Rethinking Schools*, Summer, 2001; "Standards and the Control of Knowledge" *Rethinking Schools*, Spring, 1999; "Cultural Politics, the Science of Assessment and Democratic Renewal of Public Education" in Ann Filer, ed. *Assessment: Social Practice and Social Product*, Routledge/Falmer, 2000. He lives in Oakland, and can be reached at hberlak@infinex.com.